CLAIMS

1. A method comprising:

receiving a frame of content;

automatically detecting a candidate area for a new face region in the frame;

using one or more hierarchical verification levels to verify whether a human face is in the candidate area;

indicating that the candidate area includes a face if the one or more hierarchical verification levels verify that a human face is in the candidate area; and

using a plurality of cues to track each verified face in the content from frame to frame.

2. A method as recited in claim 1, wherein the frame of content comprises a frame of video content.

3. A method as recited in claim 1, wherein the frame of content comprises a frame of audio content.

4. A method as recited in claim 1, wherein the frame of content comprises a frame of both video and audio content.

5. A method as recited in claim 1, further comprising repeating the automatic detecting in the event tracking of a verified face is lost.

6. A method as recited in claim 1, wherein receiving the frame of content comprises receiving a frame of video content from a video capture device local to a system implementing the method.

7. A method as recited in claim 1, wherein receiving the frame of content comprises receiving the frame of content from a computer readable medium accessible to a system implementing the method.

8. A method as recited in claim 1, wherein detecting the candidate area for the new face region in the frame comprises:

detecting whether there is motion in the frame and, if there is motion in the frame, then performing motion-based initialization to identify one or more candidate areas;

detecting whether there is audio in the frame, and if there is audio in the frame, then performing audio-based initialization to identify one or more candidate areas; and

using, if there is neither motion nor audio in the frame, a fast face detector to identify one or more candidate areas.

9. A method as recited in claim 1, wherein detecting the candidate area for the new face region in the frame comprises:

determining whether there is motion at a plurality of pixels on a plurality of lines across the frame;

generating a sum of frame differences for each possible segment of each of the plurality of lines;

selecting, for each of the plurality of lines, the segment having the largest sum;

identifying a smoothest region of the selected segments;

checking whether the smoothest region resembles a human upper body; and

extracting, as the candidate area, the portion of the smoothest region that resembles a human head.

10. A method as recited in claim 9, wherein determining whether there is motion comprises:

determining, for each of the plurality of pixels, whether a difference between an intensity value of the pixel in the frame and an intensity value of a corresponding pixel in one or more other frames exceeds a threshold value.

11. A method as recited in claim 1, wherein the one or more hierarchical verification levels include a coarse level and a fine level, wherein the coarse level can verify whether the human face is in the candidate area faster but with less accuracy than the fine level.

12. A method as recited in claim 1, wherein using one or more hierarchical verification levels comprises, as one of the levels of verification:

generating a color histogram of the candidate area;

generating an estimated color histogram of the candidate area based on previous frames;

determining a similarity value between the color histogram and the estimated color histogram; and

verifying that the candidate area includes a face if the similarity value is greater than a threshold value.

13. A method as recited in claim 1, wherein indicating that the candidate area includes a face comprises recording the candidate area in a tracking list.

14. A method as recited in claim 13, wherein recording the candidate area in the tracking list comprises accessing a record corresponding to the candidate area and resetting a time since last verification of the candidate.

15. A method as recited in claim 1, wherein the one or more hierarchical verification levels include a first level and a second level, and wherein using the one or more hierarchical verification levels to verify whether the human face is in the candidate area comprises:

checking whether, using the first level verification, the human face is verified as in the candidate area; and

using the second level verification only if the checking indicates that the human face is not verified as in the candidate area by the first level verification.

16. A method as recited in claim 1, wherein using one or more hierarchical verification levels comprises:

using a first verification process to determine whether the human head is in the candidate area; and

if the first verification process verifies that the human head is in the candidate area, then indicating the area includes a face, and otherwise using a second verification process to determine whether the human head is in the area.

17. A method as recited in claim 16, wherein the first verification process is faster but less accurate than the second verification process.

18. A method as recited in claim 1, wherein the plurality of cues include foreground color, background color, edge intensity, motion, and audio.

19. A method as recited in claim 1, wherein using the plurality of cues to track each verified face comprises, for each face:

predicting where a contour of the face will be;

encoding a smoothness constraint that penalizes roughness;

applying the smoothness constraint to a plurality of possible contour locations; and

selecting the contour location having the smoothest contour as the location of the face in the frame.

20. A method as recited in claim 19, wherein the smoothness constraint includes contour smoothness.

21. A method as recited in claim 19, wherein the smoothness constraint includes both contour smoothness and region smoothness.

22. A method as recited in claim 19, wherein encoding the smoothness constraint comprises generating Hidden Markov Model (HMM) state transition probabilities.

23. A method as recited in claim 19, wherein encoding the smoothness constraint comprises generating Joint Probability Data Association Filter (JPDAF) state transition probabilities.

24. A method as recited in claim 19, wherein using the plurality of cues to track each verified face further comprises, for each face:

adapting the predicting for the face in subsequent frames to account for changing color distributions.

25. A method as recited in claim 19, wherein using the plurality of cues to track each verified face further comprises, for each face:

adapting the predicting for the face in subsequent frames based on one or more cues observed in the frame.

26. A method as recited in claim 1, wherein using the plurality of cues to track each verified face comprises, for each face:

accessing a set of one or more feature points of the face;

analyzing the frame to identify an area that includes the set of one or more feature points;

encoding a smoothness constraint that penalizes roughness;

applying the smoothness constraint to a plurality of possible contour locations; and

selecting the contour location having the smoothest contour as the location of the face in the frame.

27. A method as recited in claim 1, wherein using the plurality of cues to track each verified face comprises concurrently tracking multiple possible locations for the face from frame to frame.

28. A method as recited in claim 27, further comprising using a multiple-hypothesis tracking technique to concurrently track the multiple possible locations.

29. A method as recited in claim 27, further comprising using a particle filter to concurrently track the multiple possible locations.

30. A method as recited in claim 27, further comprising using an unscented particle filter to concurrently track the multiple possible locations.

31. A system to track multiple individuals in video content, the system comprising:

an auto-initialization module to detect a candidate region for a new face in a frame of the video content;

a hierarchical verification module to generate a confidence level for the candidate region; and

a multi-cue tracking module to use a plurality of visual cues to track previous candidate regions with confidence levels, generated by the hierarchical verification module, that exceeded a threshold value.

32. A system as recited in claim 31, wherein the hierarchical verification module is further configured to:

check whether the confidence level exceeds the threshold value;

if the confidence level does exceed the threshold value then to pass the candidate region to the multi-cue tracking module; and

if the confidence level does not exceed the threshold value then to discard the candidate region and not pass the candidate region to the multi-cue tracking module.

33. A system as recited in claim 31, wherein the hierarchical verification module is further configured to:

receive, from the multi-cue tracking module, an indication of a region;

verify whether the region is a face; and

return the region to the multi-cue tracking module for continued tracking only if the region is verified as a face.

34. A system as recited in claim 31, wherein the system comprises a video conferencing system.

35. A system as recited in claim 31, wherein the auto-initialization module is further to:

detect whether there is motion in the frame;

if there is motion in the frame, then perform motion-based initialization to identify the candidate region;

detect whether there is audio in the frame;

if there is audio in the frame, then perform audio-based initialization to identify the candidate region; and

if there is neither motion in the frame nor audio in the frame, then use a fast face detector to identify the candidate region.

36. A system as recited in claim 31, wherein the hierarchical verification module is to use one or more hierarchical verification levels that include a coarse level and a fine level, wherein the coarse level can verify whether the new face is in the candidate area faster but with less accuracy than the fine level.

37. One or more computer readable media having stored thereon a plurality of instructions that, when executed by one or more processors, causes the one or more processors to:

receive an indication of an area of a frame of video content;

use a first verification process to determine whether a human head is in the area; and

if the first verification process verifies that the human head is in the area, then indicate the area includes a face, and otherwise use a second verification process to determine whether the human head is in the area.

38. One or more computer readable media as recited in claim 37, wherein the first verification process and the second verification process correspond to a plurality of hierarchical verification levels.

39. One or more computer readable media as recited in claim 38, wherein the plurality of hierarchical verification levels comprise more than two hierarchical verification levels.

40. One or more computer readable media as recited in claim 37, wherein the first verification process is a coarse level process and the second verification process is a fine level process, and wherein the coarse level process can verify whether the human head is in the candidate area faster but with less accuracy than the fine level process.

41. One or more computer readable media as recited in claim 37, wherein the plurality of instructions to use the first verification process comprises instructions that cause the one or more processors to:

generate a color histogram of the area;

generate an estimated color histogram of the area based on previous frames of the video content;

determine a similarity value between the color histogram and the estimated color histogram; and

verify that the candidate area includes the human head if the similarity value is greater than a threshold value.

42. One or more computer readable media as recited in claim 37, wherein the plurality of instructions to receive the indication of the area of the frame of video content comprises instructions that cause the one or more processors to:

receive a candidate area for a new face region in the frame.

43. One or more computer readable media as recited in claim 37, wherein the plurality of instructions to receive the indication of the area of the frame of video content comprises instructions that cause the one or more processors to:

receive an indication of an area to re-verify as including a face.

44. One or more computer readable media having stored thereon a plurality of instructions to detect a candidate region for an untracked face in a frame of content, wherein the plurality of instructions, when executed by one or more processors, causes the one or more processors to:

detect whether there is motion in the frame;

if there is motion in the frame, then perform motion-based initialization to identify the candidate region;

detect whether there is audio in the frame;

if there is audio in the frame, then perform audio-based initialization to identify the candidate region; and

if there is neither motion in the frame nor audio in the frame, then use a fast face detector to identify the candidate region.

45. One or more computer readable media as recited in claim 44, wherein the plurality of instructions to perform motion-based initialization comprises instructions that cause the one or more processors to:

determine whether there is motion at a plurality of pixels on a plurality of lines across the frame;

generate a sum of frame differences for a plurality of segments of multiple ones of the plurality of lines;

select, for each of the multiple lines, the segment having the largest sum;

identify a smoothest region of the selected segments;

check whether the smoothest region resembles a human upper body; and

extract, as the candidate area, the portion of the smoothest region that resembles a human head.

46. One or more computer readable media as recited in claim 45, wherein the instructions to determine whether there is motion comprise instructions that cause the one or more processors to:

determine, for each of the plurality of pixels, whether a difference between an intensity value of the pixel in the frame and an intensity value of a corresponding pixel in one or more other frames exceeds a threshold value.

47. One or more computer readable media having stored thereon a plurality of instructions to track faces from frame to frame of content, wherein the plurality of instructions, when executed by one or more processors, causes the one or more processors to:

predict, using a plurality of cues, where a contour of a face will be in a frame;

encode a smoothness constraint that penalizes roughness;

apply the smoothness constraint to a plurality of possible contour locations; and

select the contour location having the smoothest contour as the location of the face in the frame.

48. One or more computer readable media as recited in claim 47, wherein the plurality of cues include foreground color, background color, edge intensity, and motion.

49. One or more computer readable media as recited in claim 47, wherein the plurality of cues include audio.

50. One or more computer readable media as recited in claim 47, wherein the smoothness constraint includes contour smoothness.

51. One or more computer readable media as recited in claim 47, wherein the smoothness constraint includes both contour smoothness and region smoothness.

52. One or more computer readable media as recited in claim 47, wherein the plurality of instructions to encode the smoothness constraint comprises instructions that cause the one or more processors to generate Hidden Markov Model (HMM) state transition probabilities.

53. One or more computer readable media as recited in claim 47, wherein the plurality of instructions to encode the smoothness constraint comprises instructions that cause the one or more processors to generate Joint Probability Data Association Filter (JPDAF) state transition probabilities.

54. One or more computer readable media as recited in claim 47, wherein the plurality of instructions further comprise instructions that cause the one or more processors to:

adapt the predicting for the face in subsequent frames to account for changing color distributions.

55. One or more computer readable media as recited in claim 47, wherein the plurality of instructions further comprise instructions that cause the one or more processors to:

adapt the predicting for the face in subsequent frames based on one or more cues observed in the frame.

56. One or more computer readable media as recited in claim 47, the plurality of instructions further comprise instructions that cause the one or more processors to concurrently track multiple possible locations for the face from frame to frame.

57. One or more computer readable media as recited in claim 56, the plurality of instructions further comprise instructions that cause the one or more processors to concurrently track the multiple possible locations.

58. A method for tracking an object along frames of content, the method comprising:

using a plurality of cues to track the object.

59. A method as recited in claim 58, wherein the plurality of cues include foreground color, background color, edge intensity, motion, and audio.

60. A method as recited in claim 58, wherein the using comprises predicting wherein the object will be from frame to frame based on the plurality of cues.

61. A method for tracking an object along frames of content, the method comprising:

predicting where the object will be in a frame;

encoding a smoothness constraint that penalizes roughness;

applying the smoothness constraint to a plurality of possible object locations; and

selecting the object location having the smoothest contour as the location of the object in the frame.

62. A method as recited in claim 61, wherein the predicting uses a plurality of cues that include foreground color, background color, edge intensity, motion, and audio.

63. A method as recited in claim 61, wherein the smoothness constraint includes both contour smoothness and region smoothness.

64. A method as recited in claim 61, wherein encoding the smoothness constraint comprises generating Hidden Markov Model (HMM) state transition probabilities.

65. A method as recited in claim 61, wherein encoding the smoothness constraint comprises generating Joint Probability Data Association Filter (JPDAF) state transition probabilities.

66. A method as recited in claim 61, wherein using the plurality of cues to track each verified face further comprises, for each face:

adapting the predicting for the face in subsequent frames based on one or more cues observed in the frame.

67. A method as recited in claim 61, wherein predicting where the object will be comprises:

accessing a set of one or more feature points of the face; and

analyzing the frame to identify an area that includes the set of one or more feature points.

68. A method as recited in claim 61, wherein using the plurality of cues to track each verified face comprises concurrently tracking multiple possible locations for the face from frame to frame.

69. A method as recited in claim 68, further comprising using a multiple-hypothesis tracking technique to concurrently track the multiple possible locations.

70. A method as recited in claim 61, wherein the object comprises a face in video content.

71. A method as recited in claim 61, wherein the object comprises a sound source location in audio content.